# Notes on Statistical Inference

## ASTP 611-01: Statistical Methods for Astrophysics[*]

Spring Semester 2014

## Contents

[*] Copyright 2014, John T. Whelan, and all that

**Thursday, March 20, 2014**

# 1 Methods of Inference

Our studies of probability theory have primarily shown us how to predict the outcome of experiments given some model and/or set of parameters, to calculate $P(D|H,I)$ where $D$ represents the data, $H$ the hypothesis (possibly including parameter values) and $I$ represents any other background information. The goal of statistical inference is to take the outcome of an experiment, and say something about the validity of one or more hypotheses.

From the Bayesian point of view, this is as simple as using Bayes's Theorem to construct

$$P(H|D,I) = \frac{P(D|H,I)P(H|I)}{P(D|I)} \tag{1.1}$$

In the frequentist approach, we're not allowed to assign probabilities to hypotheses, so instead we have to use $P(D|H,I)$ to say something about the hypothesis $H$ once we know the value of $D$. In practice, this often involves dividing up the space of possible values of $D$ into a region $\mathcal{D}$ which is in some sense "consistent" with $H$, i.e., likely given $H$, so that $\sum_{D\in\mathcal{D}} P(D|H,I)$ is above some threshold. But it's a bit arbitrary to choose such regions. After all, even if a coin is fair, the exact sequence HTTTTHHTHTHTT is very unlikely (one in $2^{13}$), but it's somehow more consistent with a fair coin than a string of 12 heads and a tail would be. So we often find ourselves constructing a statistic, a single function of the data which we can use for a simple threshold. So for example, to check if a coin is fair, given that we flipped $k$ heads in $n$ tries, we could take $(k-n/2)^2$. If this is small, it the number of heads is close to what we'd expect from a fair coin. This is a goodness-of-fit statistic, and the chi-square statistics we've constructed so far are examples.

Another example of a statistic would be if we want to estimate the probability parameter associated with a binomial distribution. Given $k$ successes in $n$ trials, we'd expect $k/n$ to be a sensible estimate of this parameter. This discards any information about the order of successes and failures, and just retains that one number.

## 1.1 Statistics Constructed from Data: Two Approaches

To see how a preferred statistic might arise, let's consider the case where we have $n$ data points $\{x_i\}$, drawn from independent distributions with the same unknown mean $\mu$ and different unknown variances $\{\sigma_i^2\}$. We are thus basically making $n$ independent measurements of some unknown quantity $\mu$, each with its own error of standard deviation $\sigma_i$. How can we use the values $\{x_i\}$ to say something about $\mu$?

### 1.1.1 Bayesian Approach: Posterior pdf

The Bayesian answer to that question is straightforward: construct the posterior pdf

$$f(\mu|\{x_i\},\{\sigma_i\},I) = \frac{f(\{x_i\}|\mu,\{\sigma_i\},I)f(\mu|\{\sigma_i\},I)}{f(\{x_i\}|\{\sigma_i\},I)} \tag{1.2}$$

(From here on, we'll suppress the implicit conditional dependence on $\{\sigma_i\}$ and $I$ in the interest of compactness of notation.) To do this construction, we need to know the form of the joint pdf

$$f(\{x_i\}|\mu) = \prod_{i=1}^{n} f(x_i|\mu) \tag{1.3}$$

so let's add the additional assumption that the errors are Gaussian, so

$$f(x_i|\mu) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma_i^2}\right) \tag{1.4}$$

and

$$\begin{aligned} f(\{x_i\}|\mu) &= \prod_{i=1}^{n} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma_i^2}\right) \\ &= \frac{1}{(2\pi)^{n/2}\prod_{i=1}^{n}\sigma_i} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma_i^2}\right) \end{aligned} \tag{1.5}$$

Although this is a pdf for the $\{x_i\}$, when we substitute this likelihood function into (1.2), we will end up with a pdf for $\mu$, so we're most interested in the $\mu$ dependence, which we can see is Gaussian, since the sum in the exponential is quadratic in $\mu$. We can write this in a transparent way by completing the square and writing

$$\chi^2(\{x_i\};\mu) = \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma_i^2} = \frac{[\mu-\mu_0(\{x_i\})]^2}{\sigma_\mu^2(\{x_i\})} + \chi_0^2(\{x_i\}) \tag{1.6}$$

and solving for $\mu_0(\{x_i\})$, $\sigma_\mu^2(\{x_i\})$, and $\chi_0^2(\{x_i\})$ (the names of which have been deliberately somewhat provocatively chosen). Expanding both sides gives us

$$\begin{aligned} &\mu^2\sum_{i=1}^{n}\frac{1}{\sigma_i^2} - 2\mu\sum_{i=1}^{n}\frac{x_i}{\sigma_i^2} + \sum_{i=1}^{n}\frac{x_i^2}{\sigma_i^2} \\ &\quad = \frac{\mu^2}{\sigma_\mu^2(\{x_i\})} - 2\mu\frac{\mu_0(\{x_i\})}{\sigma_\mu^2(\{x_i\})} + \frac{[\mu_0(\{x_i\})]^2}{\sigma_\mu^2(\{x_i\})} + \chi_0^2(\{x_i\}) \end{aligned} \tag{1.7}$$

so we can solve for

$$\frac{1}{\sigma_\mu^2(\{x_i\})} = \sum_{i=1}^{n}\frac{1}{\sigma_i^2} \tag{1.8}$$

(which we see is actually independent of the data $\{x_i\}$ so we will just write $\sigma_\mu$ from now on)

$$\mu_0(\{x_i\}) = \sigma_\mu^2(\{x_i\})\sum_{i=1}^{n}\frac{x_i}{\sigma_i^2} = \frac{\sum_{i=1}^{n}\sigma_i^{-2}x_i}{\sum_{i=1}^{n}\sigma_i^{-2}} \tag{1.9}$$

and

$$\chi_0^2(\{x_i\}) = \sum_{i=1}^{n}\frac{x_i^2}{\sigma_i^2} - \frac{[\mu_0(\{x_i\})]^2}{\sigma_\mu^2(\{x_i\})} \tag{1.10}$$

While $\chi_0^2(\{x_i\})$ is useful for some applications to come later in the semester, it will turn out to be irrelevant right now, so we don't bother to work out the explicit form.

We can rewrite the likelihood function to stress its $\mu$ dependence:

$$f(\{x_i\}|\mu) = \frac{e^{-\chi_0^2(\{x_i\})/2}}{(2\pi)^{n/2}\prod_{i=1}^{n}\sigma_i} \exp\left(-\frac{[\mu-\mu_0(\{x_i\})]^2}{2\sigma_\mu^2}\right) \tag{1.11}$$

Because the posterior pdf $f(\mu|\{x_i\})$ is guaranteed to be normalized:

$$\int_{-\infty}^{\infty} f(\mu|\{x_i\})\,d\mu = \frac{\int_{-\infty}^{\infty} f(\{x_i\}|\mu)\,f(\mu)\,d\mu}{f(\{x_i\})} = 1 \tag{1.12}$$

we can write

$$f(\mu|\{x_i\}) \propto \exp\left(-\frac{[\mu-\mu_0(\{x_i\})]^2}{2\sigma_\mu^2}\right) f(\mu) \tag{1.13}$$

where the $\{x_i\}$-dependent proportionality constant can be worked out from the normalization. Explicitly,

$$f(\mu|\{x_i\}) = \mathcal{C}(\{x_i\}) \exp\left(-\frac{[\mu-\mu_0(\{x_i\})]^2}{2\sigma_\mu^2}\right) f(\mu) \tag{1.14}$$

where

$$\begin{aligned}\mathcal{C}(\{x_i\}) &= \frac{e^{-\chi_0^2(\{x_i\})/2}}{f(\{x_i\})(2\pi)^{n/2}\prod_{i=1}^n \sigma_i} \\ &= \left(\int_{-\infty}^{\infty} \exp\left(-\frac{[\mu-\mu_0(\{x_i\})]^2}{2\sigma_\mu^2}\right) f(\mu)\, d\mu\right)^{-1}\end{aligned} \tag{1.15}$$

In particular, if the prior pdf $f(\mu)$ is constant[1], the posterior is

$$f(\mu|\{x_i\}) = \frac{1}{\sigma_\mu^2\sqrt{2\pi}} \exp\left(-\frac{[\mu-\mu_0(\{x_i\})]^2}{2\sigma_\mu^2}\right) \tag{1.16}$$

In any event, no matter what the prior on $\mu$, the essential information about the outcome of the experiment is encoded in the weighted average $\mu_0(\{x_i\})$.

### 1.1.2 Frequentist Approach: Optimal Estimator

Now let's shift to the frequentist perspective, where we have $n$ random variables $\{X_i\}$ with unknown mean $\langle X_i\rangle = \mu$ and known variances $\operatorname{Cov}(X_i, X_j) = \delta_{ij}\operatorname{Var}(X_i) = \delta_{ij}\sigma_i^2$. We want to say something about $\mu$, and the simplest thing we can do is try to estimate its value. So we construct a statistic $\widehat{\mu}(\{X_i\})$. This is a random variable, and for any data realization it is our guess for the value of $\mu$. Since it's a random variable, it has an expectation value. We say that $\widehat{\mu}$ is an *unbiased estimator* of $\mu$ if $\langle\widehat{\mu}(\{X_i\})\rangle = \mu$. There are a lot of possible statistics which satisfy this requirement. For instance, we could just take $X_1$ and throw away the rest of the data. Or we could take the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Either one of these is an unbiased estimator (since they both have expectation value $\mu$), but we'd expect $\overline{X}$ to do a better job of estimating $\mu$. On the other hand, it won't be the best in all cases; for example, if $\sigma_2$ is much less than all the other $\{\sigma_i\}$, i.e., the second measurement is good and the others are all lousy, we'd like to pay more attention to $X_2$ than the other random variables.

We'd like to consider what is the best estimator $\widehat{\mu}$ to use. For simplicity, let's restrict ourselves to linear combinations of the random variables, i.e., estimators of the form

$$\widehat{\mu}(\{X_i\}) = \sum_{i=1}^n a_i X_i \tag{1.17}$$

the estimator will be unbiased if

$$\langle\widehat{\mu}(\{X_i\})\rangle = \sum_{i=1}^n a_i\mu = \mu\sum_{i=1}^n a_i \tag{1.18}$$

is equal to $\mu$, i.e., if $\sum_{i=1}^n a_i = 1$. The variance of the estimator

[1] In practice, to have a normalizable prior, we need something like

$$f(\mu) = \begin{cases}\frac{1}{\mu_{\max}-\mu_{\min}} & \mu_{\min} < \mu < \mu_{\max} \\ 0 & \text{otherwise}\end{cases}$$

but in the limit $\mu_{\min} \ll \mu_0 - \sigma$ and $\mu_{\max} \gg \mu_0 + \sigma$ we get the simpler result given here.

is

$$\mathrm{Var}\left(\widehat{\mu}(\{X_i\})\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \,\mathrm{Cov}(X_i, X_j) = \sum_{i=1}^{n} a_i^2 \sigma_i^2 \tag{1.19}$$

The *optimal estimator* is the unbiased estimator with the lowest variance, i.e., it minimizes $\sum_{i=1}^{n} a_i^2\sigma_i^2$ subject to the constraint $\sum_{i=1}^{n} a_i = 1$. We can find this with the method of Lagrange multipliers, by minimizing

$$\sum_{i=1}^{n} a_i^2\sigma_i^2 + \lambda(\sum_{i=1}^{n} a_i - 1) \tag{1.20}$$

with respect to $\{a_i\}$ and $\lambda$. Taking $\frac{\partial}{\partial a_i}$ gives

$$2a_i\sigma_i^2 + \lambda = 0 \tag{1.21}$$

so

$$a_i = -\frac{\lambda}{2}\sigma_i^{-2} \tag{1.22}$$

Taking $\frac{\partial}{\partial\lambda}$ gives the constraint $\sum_{i=1}^{n} a_i = 1$ so

$$-\frac{\lambda}{2}\sum_{i=1}^{n}\sigma_i^{-2} = 1 \tag{1.23}$$

i.e.,

$$-\frac{\lambda}{2} = \frac{1}{\sum_{i=1}^{n}\sigma_i^{-2}} \tag{1.24}$$

and

$$a_i = \frac{\sigma_i^{-2}}{\sum_{j=1}^{n}\sigma_j^{-2}} \tag{1.25}$$

which makes the optimal estimator

$$\widehat{\mu}_{\mathrm{opt}}(\{X_i\}) = \frac{\sum_{i=1}^{n}\sigma_i^{-2}X_i}{\sum_{i=1}^{n}\sigma_i^{-2}} \tag{1.26}$$

and its variance

$$\mathrm{Var}(\widehat{\mu}_{\mathrm{opt}}(\{X_i\})) = \sum_{i=1}^{n} a_i^2\sigma_i^2 = \frac{\sum_{i=1}^{n}\sigma_i^{-4}\sigma_i^2}{(\sum_{j=1}^{n}\sigma_j^{-2})^2} = \frac{1}{\sum_{i=1}^{n}\sigma_i^{-2}} \tag{1.27}$$

but we see that this optimal estimator is just the same weighted average that showed up in the posterior pdf for $\mu$ in the Bayesian approach:

$$\widehat{\mu}_{\mathrm{opt}}(\{x_i\}) = \mu_0(\{x_i\}) \tag{1.28}$$

and its variance is the width of the posterior on $\mu$ in the case where the prior is uniform and the sampling distribution is Gaussian.

$$\mathrm{Var}(\widehat{\mu}_{\mathrm{opt}}(\{X_i\})) = \sigma_\mu^2 \tag{1.29}$$

### Tuesday, April 1, 2014

## 2 Parameter Estimation

Our preceding example considered two related questions about unknown parameters

- What posterior distribution do we assign to an unknown parameter in light of observed data, in the Bayesian framework?
- How can we estimate an unknown parameter given observed data?

In addition to the Bayesian vs frequentist issues, there are also differences between trying to get a single point estimate of a parameter, and saying something about the uncertainty associated with that estimate.

## 2.1 Maximum likelihood estimates

As we saw previously, there are many different estimators that could conceivably be used to try to gain information about an unknown parameter $\theta$. One way, in the frequentist picture, to pick an estimate is the so-called maximum likelihood method, which chooses the value that maximizes the likelihood function $f(\{x_i\}|\theta)$ where $\{x_i\}$ are the observed data.

In the previous example, where the parameter was $\mu$, the likelihood function was

$$\begin{aligned} f(\{x_i\}|\mu) &= \frac{1}{(2\pi)^{n/2}\prod_{i=1}^{n}\sigma_i}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma_i^2}\right) \\ &= \frac{e^{-\chi_0^2(\{x_i\})/2}}{(2\pi)^{n/2}\prod_{i=1}^{n}\sigma_i}\exp\left(-\frac{[\mu-\mu_0(\{x_i\})]^2}{2\sigma_\mu^2}\right) \end{aligned} \tag{2.1}$$

where $\frac{1}{\sigma_\mu^2} = \sum_{i=1}^{n}\frac{1}{\sigma_i^2}$, $\mu_0(\{x_i\}) = \frac{\sum_{i=1}^{n}\sigma_i^{-2}x_i}{\sum_{i=1}^{n}\sigma_i^{-2}}$, and $\chi_0^2(\{x_i\}) = \sum_{i=1}^{n}\frac{x_i^2}{\sigma_i^2} - \frac{[\mu_0(\{x_i\})]^2}{\sigma_\mu^2(\{x_i\})}$. We can see by inspection that the likelihood function (which happens to be a Gaussian) is maximized when $\mu = \mu_0(\{x_i\})$.

As another example, consider a random sample $\{X_i\}$ of size $n$ drawn from an exponential distribution with rate parameter $\theta$. Since each random variable $X_i$ is drawn from the pdf $f(x_i|\theta) = \theta e^{\theta x_i}$, the likelihood function is

$$f(\{x_i\}|\theta) = \prod_{i=1}^{n} f(x_i|\theta) = \theta^n e^{-\theta\sum_{i=1}^{n}x_i} \tag{2.2}$$

It's actually easiest to find the $\theta$ that maximizes the likelihood by considering the log-likelihood

$$\ell(\theta) = \ln f(\{x_i\}|\theta) = n\ln\theta - \theta\sum_{i=1}^{n}x_i \tag{2.3}$$

whose derivative is

$$\ell'(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n}x_i \tag{2.4}$$

so the maximum-likelihood rate is

$$\widehat{\theta} = \frac{n}{\sum_{i=1}^{n}x_i} = \frac{1}{\overline{x}} \tag{2.5}$$

Note that in the Bayesian approach we could simply find the value of $\theta$ which maximizes $f(\theta|\{x_i\}) \propto f(\{x_i\}|\theta)f(\theta)$; if the prior $f(\theta)$ is uniform, that's the same as the maximum likelihood estimate. Note, though, that if we do a change of variables on the parameter, the maximum-likelihood point won't change, but the maximum-posterior point will. For instance, if we parametrize the exponential distribution in terms of a rate parameter $\beta = \theta^{-1}$, the likelihood function is

$$f(\{x_i\}|\beta) = \beta^{-n}e^{-\sum_{i=1}^{n}x_i/\beta} \tag{2.6}$$

and the derivative of the log-likelihood is

$$\frac{d}{d\beta}\ln f(\{x_i\}|\beta) = \frac{-n}{\beta} + \frac{\sum_{i=1}^{n}x_i}{\beta^2} \tag{2.7}$$

which is zero when

$$\beta = \frac{\sum_{i=1}^{n}x_i}{n} = \widehat{\theta}^{-1} \tag{2.8}$$

The reason this doesn't work for the maximum-posterior point is that $f(\theta|\{x_i\})$ is a density in $\theta$, while $f(\{x_i\}|\theta)$ is not. On the one hand,

$$f_{X|\mathrm{B}}(x|\beta) = f_{X|\Theta}(x|\beta^{-1}) \tag{2.9}$$

because the condition $\mathrm{B} = \beta$ is the same as the condition $\Theta = \beta^{-1}$, but if we transform the pdf,

$$f_{\mathrm{B}|X}(\beta|x) = \frac{dP}{d\beta} = \left|\frac{d\theta}{d\beta}\right|\frac{dP}{d\theta} = \beta^{-2}f_{\Theta|X}(\beta^{-1}|x) \tag{2.10}$$

(By the same token, the statement “the prior is uniform in the parameter” depends on what the parameter is. If $f_\Theta(\theta)$ is a constant, $f_\mathrm{B}(\beta) = \beta^{-2} f_\Theta(\beta^{-1})$ can't be.

### 2.1.1 The Fisher information matrix

Whether we're dealing with the likelihood function or the posterior, we can ask how this function behaves in the vicinity of the parameter value which maximizes it. Suppose, for concreteness, we're talking about the likelihood function $L(\theta) = f(\mathbf{x}|\theta)$. One trick would be to Taylor expand the function near its maximum, but this could cause trouble if we extrapolate it too far, since we know $f(\mathbf{x}|\theta) \geq 0$. So instead, we Taylor expand the logarithm $\ell(\theta) = \ln L(\theta)$ The expansion looks like

$$\ell(\theta) = \ell(\widehat{\theta}) + (\theta - \widehat{\theta})\ell'(\widehat{\theta}) + \frac{(\theta - \widehat{\theta})^2}{2}\ell''(\widehat{\theta}) + \cdots \tag{2.11}$$

Now, since $\widehat{\theta}$ maximizes $\ell(\theta)$, we know $\ell'(\widehat{\theta}) = 0$ and $\ell''(\widehat{\theta}) < 0$. If we truncate the expansion at the first non-trivial order, we have

$$\ell(\theta) \approx \ell(\widehat{\theta}) - \frac{(\theta - \widehat{\theta})^2}{2}[-\ell''(\widehat{\theta})] \tag{2.12}$$

or

$$L(\theta) \approx L(\widehat{\theta}) \exp\left(-\frac{(\theta - \widehat{\theta})^2}{2}[-\ell''(\widehat{\theta})]\right) \tag{2.13}$$

which is a Gaussian with width $[-\ell''(\widehat{\theta})]^{-1/2}$.

The second derivative $-\ell''(\widehat{\theta})$ is the one-dimensional version of what's known as the Fisher information matrix. In the case where there are multiple parameters, $\boldsymbol{\theta} \equiv \{\theta_i | i = 1, \ldots, m\}$, the Taylor expansion of $\ell(\boldsymbol{\theta}) = \ln L(\boldsymbol{\theta}) = \ln f(\mathbf{x}|\boldsymbol{\theta})$ is

$$\ell(\boldsymbol{\theta}) \approx \ell(\widehat{\boldsymbol{\theta}}) + \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \frac{\partial^2 \ell}{\partial\theta_i \partial\theta_j}(\theta_i - \widehat{\theta}_i)(\theta_j - \widehat{\theta}_j) \tag{2.14}$$

so that

$$L(\boldsymbol{\theta}) \approx L(\widehat{\boldsymbol{\theta}}) \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^\mathrm{T}\mathbf{F}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})\right) \tag{2.15}$$

where $\mathbf{F}$ is the Fisher matrix, which has elements

$$F_{ij} = -\frac{\partial^2 \ell}{\partial\theta_i \partial\theta_j} \tag{2.16}$$

The Fisher matrix gives an estimate of uncertainties of the parameters. This is actually clearer to see in the Bayesian case, where we approximate the posterior $f(\boldsymbol{\theta}|\mathbf{x})$ as a multivariate Gaussian, rather than just the likelihood $L(\boldsymbol{\theta})$. In that case, $\mathbf{F}$ is just the inverse of the variance-covariance matrix for the approximate multivariate Gaussian posterior:

$$\mathrm{Cov}(\boldsymbol{\theta}) = \left\langle(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^\mathrm{T}\right\rangle \approx \mathbf{F}^{-1} \tag{2.17}$$

In particular, the width of the marginal pdf for a particular parameter is

$$\sqrt{\mathrm{Var}(\theta_i)} = \left\langle(\theta_i - \widehat{\theta}_i)^2\right\rangle \approx \sqrt{[\mathbf{F}^{-1}]_{ii}} \tag{2.18}$$

This is one justification for the practice, used in both frequentist and Bayesian contexts, of quoting $\sqrt{[\mathbf{F}^{-1}]_{ii}}$ as the one-sigma uncertainty for the parameter $\theta_i$.

Note that if the Fisher matrix has off-diagonal elements, it's important to take the diagonal elements of the inverse Fisher matrix rather than one over the diagonal elements of the Fisher matrix, since

$$\left[\mathbf{F}^{-1}\right]_{ii} \neq \frac{1}{F_{ii}} \tag{2.19}$$

In general $(F_{ii})^{-1/2}$ will be an underestimate of the correct error $([\mathbf{F}^{-1}]_{ii})^{1/2}$, as you showed in your consideration of the bivariate Gaussian distribution on the homework.

## 2.2 Interval estimation

Beyond finding some "most likely" parameter value and describing the shape of either the likelihood function or the posterior around that value, an important task in parameter estimation is to provide an interval that we associate quantitatively with likely values of the parameter. This can be extended to a region in a multidimensional parameter space. The biggest difference between the Bayesian and frequentist versions of these intervals turns out to be the interpretation.

## Thursday, April 3, 2014

### 2.2.1 Bayesian plausible intervals

We start with the Bayesian version, which is considerably more straightforward. Given a posterior pdf $f(\theta|\mathbf{x})$, we can construct a *plausible interval* in which we think $\theta$ is likely to lie with some probability $1-\alpha$, defined by

$$P(\theta_\ell < \theta < \theta_u) = \int_{\theta_\ell}^{\theta_u} f(\theta|\mathbf{x})\, d\theta = 1-\alpha \tag{2.20}$$

So this means the area under the posterior pdf, between $\theta_\ell$ and $\theta_u$, is $1-\alpha$. This does leave the freedom to choose where the interval begins. Some convenient choices are

- A lower limit (one-sided plausible interval), so $P(\theta_\ell < \theta) = 1-\alpha$.
- An upper limit (one-sided plausible interval), so $P(\theta < \theta_u) = 1-\alpha$.
- A symmetric two-sided plausible interval, so $P(\theta < \theta_\ell) = \alpha/2 = P(\theta_u < \theta)$.
- A plausible interval centered on the mode $\widehat{\theta}$ of the posterior, so $P(\widehat{\theta} - \frac{\Delta\theta}{2} < \theta < \widehat{\theta} + \frac{\Delta\theta}{2}) = 1-\alpha$.
- The narrowest possible plausible interval, i.e., of all of the intervals with $P(\theta_\ell < \theta < \theta_u) = 1-\alpha$, pick the one that minimizes $\theta_u - \theta_\ell$. You can show that a necessary condition for this is $f(\theta_\ell|\mathbf{x}) = f(\theta_u|\mathbf{x})$.

### 2.2.2 Frequentist confidence intervals

In the frequentist picture we can't assign a probability to the statement that a particular interval contains or doesn't contain an unknown parameter. It either does or it doesn't. So instead we can define a procedure to generate an interval such that if you collect many random data sets and make such an interval from each, some fraction of those intervals will contain the true parameter value. This is known as a (frequentist) confidence interval. It's a pair of statistics $L = L(\mathbf{X})$ and $U = U(\mathbf{X})$ chosen so that the probability that the parameter $\theta$ lies between them is $1-\alpha$ (e.g., if $\alpha = 0.10$, it is 90%):[2]

$$P(L < \theta < U) = 1-\alpha \tag{2.21}$$

It's important to note that the probabilities here refer to the randomness of $L$ and $U$, and not to the unknown $\theta$. From the frequentist perspective, we can't talk about probabilities for different values of $\theta$; it has some specific value, even if it's unknown. What's random is the sample $\mathbf{X}$ and the statistics $L$ and $U$ created from it.

Given a particular realization $\mathbf{x}$ of the sample $\mathbf{X}$, we have a specific confidence interval between $\ell = L(\mathbf{x})$ and $u = U(\mathbf{x})$. Note that the probabilistic statements do not actually refer to the properties of a particular confidence interval $(\ell, u)$ but to the procedure used to construction of the confidence interval.

[2] We're implicitly considering a *two-sided* confidence interval, so we also have $P(\theta < L) = \alpha/2$ and $P(U < \theta) = \alpha/2$.

One method to construct the confidence interval is to choose a statistic $T = T(\mathbf{X};\theta)$, known as a *pivot variable*, whose probability distribution is a known function of the parameters, and construct an interval using the percentiles of the distribution

$$P(a < T(\mathbf{X};\theta) < b) = 1 - \alpha \tag{2.22}$$

By algebraically solving the inequalities $a < T(\mathbf{X};\theta)$ and $T(\mathbf{X};\theta) < b$ for $\theta$, we should be able to write

$$P(L(\mathbf{X}) < \theta < U(\mathbf{X})) = 1 - \alpha \tag{2.23}$$

Note that this construction is not unique; different choices for the pivot variable will give different confidence intervals with the same confidence.

### 2.2.3 Example: Mean of a Normal Distribution

To illustrate the pivot variable method, consider the case where $\mathbf{X}$ is a sample of size $n$ drawn from a $N(\mu,\sigma)$ distribution with both $\mu$ and $\sigma$ unknown, where we want a confidence interval on $\mu$. The pivot variable should depend on $\mu$ and $\mathbf{X}$ but not $\sigma$, so

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \tag{2.24}$$

will not work, even though we know it obeys as $N(0,1)$ distribution (because $\overline{X}$ obeys a normal distribution with $E(\overline{X}) = \mu$ and $\mathrm{Var}(\overline{X}) = \sigma/\sqrt{n}$. Fortunately, we know from Student's theorem that

$$T = \frac{\overline{X} - \mu}{\sqrt{S^2/n}} \tag{2.25}$$

obeys a $t$ distribution with $n-1$ degrees of freedom. This will work as a pivot variable, since

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{2.26}$$

depends only on the sample, and requires no knowledge of $\mu$ or $\sigma$. Having identified a pivot variable which obeys a $t$ distribution is useful not so much because we know the precise form of the pdf

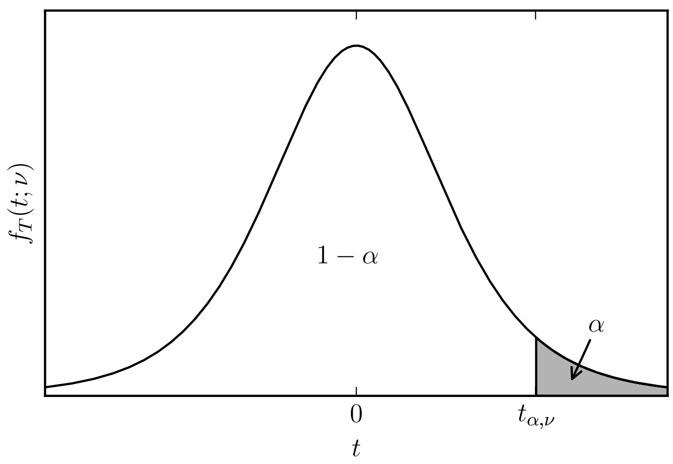
$$f_T(t;\nu) = \frac{\Gamma([\nu+1]/2)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)}\left(1 + \frac{t^2}{\nu}\right)^{-[\nu+1]/2} \tag{2.27}$$

but because it's a standard distribution for which the percentiles are tabulated in various books or available in R, scipy, etc. The 90th percentile, for example, of a $t$ distribution with $\nu$ degrees of freedom is written $t_{0.1,\nu}$; in general, the $(1-\alpha)\times 100$th percentile $t_{\alpha,\nu}$ is defined by

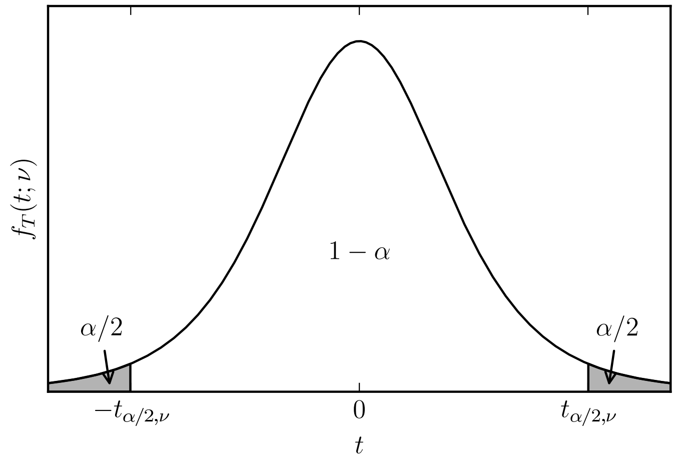
$$1 - \alpha = P(T \leq t_{\alpha,\nu}) = \int_{-\infty}^{t_{\alpha,\nu}} f_T(t;\nu)\,dt \tag{2.28}$$

or equivalently by

$$\int_{t_{\alpha,\nu}}^{\infty} f_T(t;\nu)\,dt = \alpha \tag{2.29}$$



Since we want a two-sided confidence interval, we actually need $t_{\alpha/2,\nu}$ and $t_{1-\alpha/2,\nu}$. Since the $t$ distribution is symmetric, though, we can take advantage of the fact that $t_{1-\alpha/2,\nu} = -t_{\alpha/2,\nu}$, e.g., the 5th percentile is minus the 95th:



Thus, returning to the case of the pivot variable $T$, which is $t$-distributed with $n-1$ degrees of freedom,

$$\begin{aligned} 1-\alpha &= P(-t_{\alpha/2,n-1} < T < t_{\alpha/2,n-1}) \\ &= P\left(-t_{\alpha/2,n-1} < \frac{\overline{X}-\mu}{\sqrt{S^2/n}} < t_{\alpha/2,n-1}\right) \end{aligned} \tag{2.30}$$

Doing a bit of algebra, we can see that

$$\frac{\overline{X}-\mu}{\sqrt{S^2/n}} < t_{\alpha/2,n-1} \tag{2.31}$$

is equivalent to

$$\overline{X} - t_{\alpha/2,n-1}\sqrt{\frac{S^2}{n}} < \mu \tag{2.32}$$

and

$$-t_{\alpha/2,n-1} < \frac{\overline{X}-\mu}{\sqrt{S^2/n}} \tag{2.33}$$

is equivalent to

$$\mu < \overline{X} + t_{\alpha/2,n-1}\sqrt{\frac{S^2}{n}} \tag{2.34}$$

so

$$P\left(\overline{X} - t_{\alpha/2,n-1}\sqrt{\frac{S^2}{n}} < \mu < \overline{X} + t_{\alpha/2,n-1}\sqrt{\frac{S^2}{n}}\right) = 1-\alpha \tag{2.35}$$

which defines a confidence interval for $\mu$.

## Tuesday, April 8, 2014

# 3 Model Selection

## 3.1 Frequentist hypothesis testing

*See Gregory, Chapter 7*

Often want to evaluate hypothesis $\mathcal{H}$ in light of observed data $\mathbf{x}$, or compare hypotheses

In Bayesian picture, can define $P(\mathcal{H}|\mathbf{x})$ and evaluate e.g., $\frac{P(\mathcal{H}_1)}{P(\mathcal{H}_2)}$

So far, we've considered methods to get a handle on the unknown parameter(s) $\theta$ of a probability distribution $f(x;\theta)$ given that we draw a sample $\mathbf{X}$ from that distribution, with joint pdf

$$f_{\mathbf{X}}(x_1,x_2,\ldots,x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta) \tag{3.1}$$

and find a particular realization $\mathbf{X}=\mathbf{x}$. Now we want to consider how to use the realization of the sample to distinguish between two competing hypotheses about what the underlying distribution $f(x)$ is. In principle the differences could be qualitative, but for simplicity we'll assume that there is one family $f(x;\theta)$ parametrized by $\theta$ which lies somewhere in a region $\Omega$ and then take the hypotheses to be:

- $H_0$: the distribution is $f(x;\theta)$ where $\theta \in \omega_0$.
- $H_1$: the distribution is $f(x;\theta)$ where $\theta \in \omega_1$.

Typically, $H_0$ represents the absence of the effect we're looking for, and is known as the *null hypothesis*, while $H_1$ represents the presence of the effect, and is known as the *alternative hypothesis*.

For example, suppose someone claims to have extrasensory perception, and to be able to use their telepathic powers to determine the suits of cards drawn from a deck. For simplicity, assume we shuffle the deck after each draw. Then the data $\{X_i\}$ are a sample drawn from a Bernoulli distribution, with each $X_i$ having some probability $\theta$ of being correct. The null hypothesis $H_0$ is that the person does not have ESP, and has a 25% chance of guessing each suit correctly, so $\theta = 0.25$. The alternative hypothesis $H_1$ is that they can determine the suit more accurately than by random chance (but perhaps not perfectly), so $\theta > 0.25$.

As another example, suppose that someone claims that when twins are born, the birth weight of the first twin is on average greater than that of the second. We could take the data $\{X_i\}$ to be the difference between the birth weights of the two twins, and assume that the weights are normally distributed with unknown variance. Then the null hypothesis $H_0$ is that $f(x)$ is a normal distribution with mean $\mu = 0$ and standard deviation $\sigma > 0$, while the alternative hypothesis $H_1$ is that $f(x)$ is a normal distribution with mean $\mu > 0$ and standard deviation $\sigma > 0$. (In this case there is a vector of parameters $\boldsymbol{\theta} = (\mu, \sigma)$.

A hypothesis test is simply a rule for choosing between the two hypotheses depending on the realization $\mathbf{x}$ of the sample $\mathbf{X}$. Stated most generally, we construct a critical region $C$ which is a subset of the $n$-dimensional sample space $\mathcal{D}$. If $\mathbf{X} \in C$, we "reject the null hypothesis $H_0$", i.e., we favor $H_1$. If $\mathbf{X} \notin C$, i.e., $\mathbf{X} \in C^c$ we "accept the null hypothesis $H_0$", i.e., we favor $H_0$ over $H_1$. Now of course, since $\mathbf{X}$ is random, there will be some probability $P(\mathbf{X} \in C; \theta)$ that we'll reject the null hypothesis, which depends on the value of $\theta$. If the test were perfect, that probability would be 0 if $H_0$ were true, i.e., for any $\theta \in \omega_0$, and 1 if $H_1$ were true, i.e., for any $\theta \in \omega_1$, but then we wouldn't be doing statistics. So instead there is some chance we will choose the "wrong" hypothesis, i.e., some probability that, given a value of $\theta \in \omega_0$ associated with $H_0$, the realization of our data will cause us to reject $H_0$, and some probability that, given a value of $\theta \in \omega_1$ associated with $H_1$, the realization of our data will cause us to accept $H_0$. As a bit of nomenclature,

- If $H_0$ is true and we reject $H_0$, this is called a *Type I Error* or a false positive.
- If $H_1$ is true and we reject $H_0$, we have made a correct decision (true positive).
- If $H_0$ is true and we accept $H_0$, we have made a correct decision (true negative).
- If $H_1$ is true and we accept $H_0$, this is called a *Type II Error* or a false negative.

Typically, a false positive is considered worse than a false negative, so usually we decide how high a false positive probability we can live with and then try to find the test which gives us the lowest false negative probability.

Given a critical region $C$, we'd like to talk about the associated false positive probability $\alpha$ and false negative probability $1 - \gamma$, but we have to be a bit careful, since $H_0$ and $H_1$ are in general *composite hypotheses*. This means that each of them corresponds not to a single parameter value $\theta$ and thus a single distribution, but rather to a range of values $\theta \in \omega_0$ or $\theta \in \omega_1$. So both $\alpha$ and $\gamma$ may depend on the value of $\theta$. We take the false alarm probability $\alpha$ to be the worst-case scenario within the null hypothesis

$$\alpha = \max_{\theta \in \omega_0} P(\mathbf{X} \in C; \theta) \tag{3.2}$$

This is also called the *size* of the critical region $C$. Somewhat confusingly, it's also referred to as the *significance* of the test. This is a bit counter intuitive, since a low value of $\alpha$ means the probability of a false positive is low, which means a positive result is *more* significant than if $\alpha$ were higher. It is the probability that we'll falsely reject the null hypothesis $H_0$, maximized over any parameters within the range associated with $H_0$. On the other hand, since the alternative hypothesis almost always has a parameter $\theta$ associated with it, we define the probability of correctly rejecting the null hypothesis (which is one minus the probability of a false negative) as a function of $\theta$:

$$\gamma_C(\theta) = P(\mathbf{X} \in C; \theta), \qquad \theta \in \omega_1 \tag{3.3}$$

We explicitly consider this as a function of the critical region $C$, since we might want to compare different tests with the same false alarm probability $\alpha$ (critical regions with the same size $\alpha$) to see which is more powerful.

## 3.2 Example: Binomial Proportion

To give a concrete example, consider the ESP test described above. We let the would-be psychic predict the suit of $n$ cards, count the total number of successes $Y = \sum_{i=1}^{n} X_i$, and reject the null hypothesis if $Y > k$ where $k$ is some integer we've chosen, with $k > n/4$. For both of the hypotheses, $Y$ is a binomial random variable, so

$$P(Y > k) = \sum_{i=k+1}^{n} \binom{n}{i} \theta^i (1-\theta)^{n-i} = 1 - F(k;\theta) \tag{3.4}$$

where

$$F(k;\theta) = \sum_{i=0}^{k} \binom{n}{i} \theta^i (1-\theta)^{n-i} \tag{3.5}$$

is the cdf of a binomial distribution $b(n,\theta)$. For the null hypothesis $\theta = 0.25$ and for the alternative hypothesis $0.25 < \theta < 1$. Thus the false alarm probability is

$$\alpha = 1 - F(k; 0.25) \tag{3.6}$$

and the power of the test is

$$\gamma_k(\theta) = 1 - F(k;\theta) \tag{3.7}$$

If we make the threshold $k$ higher, we get a lower false alarm probability $\alpha$, but we also get a less powerful test.

As a concrete example, suppose that $n = 20$, and we set a threshold of $k = 8$. We can use scipy, invoked by

```
ipython --pylab
```

to calculate the false alarm probability

```
In [1]: from scipy.stats import binom

In [2]: n = 20

In [3]: k = 8

In [4]: alpha = 1 - binom.cdf(k,n,0.25); alpha
Out[4]: 0.04092516770651855
```

So $\alpha \approx 0.041 = 4.1\%$. The power $\gamma(\theta)$ depends on the strength of the ESP effect, but suppose $\theta = 0.50$, that the psychic has a 1 in 2 chance rather than 1 in 4 of picking the right suit. Then we can calculate the power:

```
In [5]: gamma_50 = 1 - binom.cdf(k,n,0.50); gamma_50
Out[5]: 0.74827766418457031
```

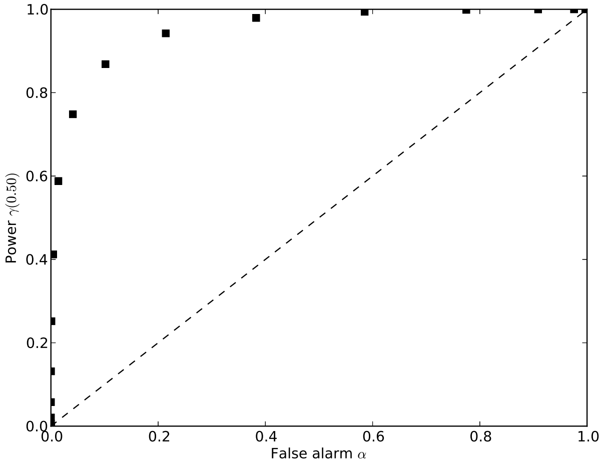
so $\gamma(0.50) \approx 0.748 = 74.8\%$.

### 3.2.1 Aside: ROC Curves

We could make the test more powerful by lowering the threshold $k$, but then we would also increase the false alarm probability $\alpha$. A useful construction is the *receiver operating characteristic* curve, or ROC curve for short. Given a value of $\theta$, we plot $\alpha$ versus $\gamma(\theta)$ for a range of threshold values $k$. We can do this with matplotlib as well, using the `arange` function to generate an array of integer values for $k$ between 0 and 19:

```
In [6]: k = arange(20)

In [7]: alpha = 1 - binom.cdf(k,n,0.25)

In [8]: gamma_50 = 1 - binom.cdf(k,n,0.50)

In [9]: plot(alpha,gamma_50,'ks');

In [10]: xlabel(r'False alarm $\alpha$');

In [11]: ylabel(r'Power $\gamma(0.50)$');

In [12]: plot([0,1],[0,1],'k--');

In [13]: savefig('roc.eps');
```

The plot looks like this:



The diagonal line is $\gamma = \alpha$; we don't expect any sensible test to lie below this line, since it would mean that we were more likely to reject $H_0$ when it's true than when $H_1$ is true!

## 3.3 Example: Mean of a Normal Distribution

Consider the second example, where $\mathbf{X}$ is a random sample of size $n$ from a normal distribution, where the null hypothesis $H_0$ is $\mu = 0$ and the alternative hypothesis $H_1$ is $\mu > 0$. For simplicity, let's assume that the variance $\sigma^2$ is actually known. (If the sample is large enough, we can use the sample variance $s^2$ as an estimate.) From our work on confidence intervals, we know that

$$P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} > z_\alpha\right) = \alpha \tag{3.8}$$

So if we define a critical region

$$C \equiv \frac{\overline{X}}{\sigma/\sqrt{n}} > z_\alpha \tag{3.9}$$

this will correspond to a test with false alarm rate $\alpha$. The power of the test for a given true value of $\mu$ is

$$\begin{aligned}\gamma(\mu) &= P\left(\frac{\overline{X}}{\sigma/\sqrt{n}} > z_\alpha\right) = P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} > z_\alpha - \frac{\mu}{\sigma/\sqrt{n}}\right) \\ &= 1 - \Phi\left(z_\alpha - \frac{\mu}{\sigma/\sqrt{n}}\right) = \Phi\left(\frac{\mu}{\sigma/\sqrt{n}} - z_\alpha\right)\end{aligned} \tag{3.10}$$

### 3.3.1 $p$-Values

In this example, as in the last one, we actually have a family of tests, parametrized by a threshold which we could imagine varying. Given a data realization $\mathbf{x}$, and in particular a sample mean $\overline{x}$, we will reject $H_0$ if $\overline{x} > z_\alpha\sigma/\sqrt{n}$. This means there will be some values of the false alarm probability $\alpha$ for which we reject $H_0$, and some for which we do not. One convenient way to report which tests would indicate a positive result (reject the null hypothesis) is to quote the $\alpha$ of the most stringent test for which $H_0$ would be rejected. Put another way, we ask, given a measurement (in this case $\overline{x}$), how likely is it that we would find a measurement at least this extreme, just by accident, if the null hypothesis were true. This is known as the $p$-value, and in this case it is defined as

$$p = P(\overline{X} \geq \overline{x}; \mu = 0) = 1 - \Phi\left(\frac{\overline{x}}{\sigma/\sqrt{n}}\right) = \Phi\left(-\frac{\overline{x}}{\sigma/\sqrt{n}}\right) \tag{3.11}$$

A lower $p$ value means that the results were less likely to have occurred by chance in the absence of a real effect (i.e., if the null hypothesis $H_0$ were true). Typically if $p < 0.05$, the result is considered interesting and worth future study.[3]

Note that the $p$ value is often misinterpreted. It does *not* represent the probability that the null hypothesis is true (we cannot evaluate such a probability in frequentist inference). A $p$ value of 0.01 simply means, for the statistic we decided to measure, if we repeated the test on many systems for which the null hypothesis was true, we'd get a measurement as extreme, or more, as the one we got, one percent of the time.

# Thursday, April 10, 2014

## 3.4 Odds ratio and Bayes factor

*See Gregory, Section 3.5 and Sivia, Chapter 4*

One of the problems about using a frequentist test like a chi-squared test to assess the validity of a model is that you can always make the fit better by adding more parameters to the model. In the extreme case, if you have as many model parameters as data points, you can make the fit perfect. But clearly a model which is "overtuned" in this way is scientifically unsatisfying.

Bayesian statistics offers a natural way to compare models, which automatically penalizes models that use too many parameters to fine-tune themselves to match a data set. This is known as the odds ratio.

Consider Bayes's theorem in the context of a model $\mathcal{M}$ with parameters $\boldsymbol{\theta}$. Given an observation $\mathbf{x}$, we can construct the posterior pdf for the parameters $\boldsymbol{\theta}$ as follows

$$f(\boldsymbol{\theta}|\mathbf{x},\mathcal{M}) = \frac{f(\mathbf{x}|\boldsymbol{\theta},\mathcal{M})f(\boldsymbol{\theta}|\mathcal{M})}{f(\mathbf{x}|\mathcal{M})} \tag{3.12}$$

which is sometimes abbreviated as

$$(\text{posterior}) = \frac{(\text{likelihood})(\text{prior})}{(\text{evidence})} \tag{3.13}$$

[3] However, if we test for many different effects, or test many different data sets, and only report the result with the lowest $p$ value, we can greatly overstate the significance of our results. See `http://xkcd.com/882/`.

So far we've just treated the denominator as a normalization factor

$$f(\mathbf{x}|\mathcal{M}) = \int d\theta\, f(\mathbf{x}|\boldsymbol{\theta},\mathcal{M}) f(\boldsymbol{\theta}|\mathcal{M}) \tag{3.14}$$

but we will now see how it gets the name "evidence". Note that it is the overall probability to get the observed result $\mathbf{x}$ given the model $\mathcal{M}$, marginalizing over the parameters $\boldsymbol{\theta}$.

Now, consider the case where $\mathcal{M}$ is one of a number of possible models, and we'd like to construct a posterior probability $P(\mathcal{M}|\mathbf{x})$ that $\mathcal{M}$ is the correct model. Well, since we have a way to calculate $f(\mathbf{x}|\mathcal{M})$, we can try using Bayes's theorem:

$$P(\mathcal{M}|\mathbf{x}) = \frac{f(\mathbf{x}|\mathcal{M})P(\mathcal{M})}{f(\mathbf{x})} \tag{3.15}$$

The right-hand side has a couple of things that are harder to get a handle on: the prior probability $P(\mathcal{M})$ of $\mathcal{M}$ being the correct model, and the overall pdf $f(\mathbf{x})$ which requires somehow marginalizing over all possible models. The usual way around this is to consider two competing models $\mathcal{M}_1$ and $\mathcal{M}_2$, and to calculate the ratio of their posteriors, known as the odds ratio

$$\begin{aligned}\mathcal{O}_{12} &= \frac{P(\mathcal{M}_1|\mathbf{x})}{P(\mathcal{M}_2|\mathbf{x})} = \frac{f(\mathbf{x}|\mathcal{M}_1)P(\mathcal{M}_1)}{f(\mathbf{x}|\mathcal{M}_2)P(\mathcal{M}_2)} = \left(\frac{f(\mathbf{x}|\mathcal{M}_1)}{f(\mathbf{x}|\mathcal{M}_2)}\right)\left(\frac{P(\mathcal{M}_1)}{P(\mathcal{M}_2)}\right) \\ &= \left(\frac{P(\mathcal{M}_1)}{P(\mathcal{M}_2)}\right)\mathcal{B}_{12}\end{aligned} \tag{3.16}$$

So the factor of $f(\mathbf{x})$ has cancelled out, and the odds ratio $\mathcal{O}_{12}$ is the ratio of prior probabilities for each model times something known as the *Bayes factor*

$$\mathcal{B}_{12} = \frac{f(\mathbf{x}|\mathcal{M}_1)}{f(\mathbf{x}|\mathcal{M}_2)} \tag{3.17}$$

which is the ratio of the "evidence" in each of the models. It represents how our relative confidence in the two probabilities has changed with the measurement $\mathbf{x}$. If each model has some parameters, the Bayes factor can be written as

$$\mathcal{B}_{12} = \frac{\int d\theta_1\, f(\mathbf{x}|\boldsymbol{\theta}_1,\mathcal{M}_1)\, f(\boldsymbol{\theta}_1|\mathcal{M}_1)}{\int d\theta_2\, f(\mathbf{x}|\boldsymbol{\theta}_2,\mathcal{M}_2)\, f(\boldsymbol{\theta}_2|\mathcal{M}_2)} \tag{3.18}$$

To see how the Bayes factor penalizes modes for over-tuning, consider a simple case where there are two models: $\mathcal{M}_0$, which has no parameters and $\mathcal{M}_1$, which has a parameter $\theta$. If we measure data $\mathbf{x}$, the Bayes factor comparing the two models is

$$\mathcal{B}_{10} = \frac{\int_{-\infty}^{\infty} d\theta\, f(\mathbf{x}|\theta,\mathcal{M}_1)\, f(\theta|\mathcal{M}_1)}{f(\mathbf{x}|\mathcal{M}_0)} \tag{3.19}$$

To get a handle on what the marginalization of the parameter $\theta$ does, as compared with the maximization done by the frequentist method, let's make some simplifying assumptions. First let's assume the likelihood $f(\mathbf{x}|\theta,\mathcal{M}_1)$, seen as a function of $\theta$, can be approximated as a Gaussian about the maximum likelihood value $\widehat{\theta}$:

$$f(\mathbf{x}|\theta,\mathcal{M}_1) \approx f(\mathbf{x}|\widehat{\theta},\mathcal{M}_1)\, e^{-(\theta-\widehat{\theta})/2\sigma_\theta^2} \tag{3.20}$$

We'll also assume that this is sharply peaked compared to the prior $f(\theta|\mathcal{M}_1)$ and therefore we can replace $\theta$ in the argument of the prior with $\widehat{\theta}$, and

$$\begin{aligned}\int_{-\infty}^{\infty} d\theta\, f(\mathbf{x}|\theta,\mathcal{M}_1)\, f(\theta|\mathcal{M}_1) &\approx f(\mathbf{x}|\widehat{\theta},\mathcal{M}_1)\, f(\widehat{\theta}|\mathcal{M}_1) \int_{-\infty}^{\infty} d\theta\, e^{-(\theta-\widehat{\theta})/2\sigma_\theta^2} \\ &= f(\mathbf{x}|\widehat{\theta},\mathcal{M}_1)\, f(\widehat{\theta}|\mathcal{M}_1)\, \sigma_\theta\sqrt{2\pi}\end{aligned} \tag{3.21}$$

We can then approximate the Bayes factor as

$$\mathcal{B}_{10} = \frac{f(\mathbf{x}|\widehat{\theta},\mathcal{M}_1)}{f(\mathbf{x}|\mathcal{M}_0)} \frac{\sigma_\theta\sqrt{2\pi}}{[f(\widehat{\theta}|\mathcal{M}_1)]^{-1}} \tag{3.22}$$

The first factor is the ratio of the likelihoods between the best-fit version of model $\mathcal{M}_1$ and the parameter-free model $\mathcal{M}_0$. That's basically the end of the story in frequentist model comparison, and we can see that if $\mathcal{M}_0$ is included as a special case of $\mathcal{M}_1$, this ratio will always be greater or equal to one, i.e., the tunable model will always be able to find a higher likelihood than the model without that tunable parameter. But in Bayesian model comparison, there is also the second factor:

$$\frac{\sigma_\theta\sqrt{2\pi}}{[f(\widehat{\theta}|\mathcal{M}_1)]^{-1}} \qquad \text{“Occam factor”} \tag{3.23}$$

This is called the *Occam factor* because it implements Occam's razor, the principle that, all else being equal, simpler explanations will be favored over more complicated ones. Because the prior $f(\theta|\mathcal{M}_1)$ is normalized, $[f(\widehat{\theta}|\mathcal{M}_1)]^{-1}$ is a measure of the width of the prior, i.e., how much parameter space the tunable model has available to it. In particular, if the prior is uniform over some range:

$$f(\theta|\mathcal{M}_1) = \begin{cases} \frac{1}{\theta_{\text{max}}-\theta_{\text{min}}} & \theta_{\text{min}} < \theta < \theta_{\text{max}} \\ 0 & \text{otherwise} \end{cases} \tag{3.24}$$

then the Occam factor becomes

$$\frac{\sigma_\theta\sqrt{2\pi}}{\theta_{\text{max}}-\theta_{\text{min}}} \tag{3.25}$$

because we assumed the likelihood function was narrowly peaked compared to the prior, the Occam factor is always less than one, and the tunable model must have a large enough increase in likelihood over the simpler model in order to overcome this.

**Tuesday, April 15, 2014**

# 4 Advanced Topic: Monte Carlo Methods

Monte Carlo, in general, refers to calculations carried out with random or pseudo-random elements. (The name refers to the Monte Carlo Casino in Monaco.) There are a number of different such methods, but we'll focus on two:

1. Monte Carlo simulations to test the validity of a statistical method
2. Markov Chain Monte Carlo (MCMC) procedures to estimate the results of high-dimensional integrals

## 4.1 Monte Carlo Simulations

1. For each iteration:
   (a) Randomly generate values using assumed distributions
   (b) Calculate statistics of interest
2. Histogram statistics (e.g., on what fraction of iterations is the $\chi^2$ above some threshold)
3. Compare observed frequencies of statistic values to predicted distribution of statistic

Note: how to simulate a random variable with a specified pdf $f(x)$:

1. Easy way/cheating: use statistical package, e.g., `scipy.stats.norm(loc=mu,scale=sigma,size=n)`
2. General approach: given cdf $F(x) = \int_{-\infty}^{x} f(x')\,dx'$, invert to get $x = F^{-1}(P)$ for $0 < P < 1$. Generate uniform random number $\alpha$ and take $x = F^{-1}(\alpha)$.

For multivariate distributions, things may be more complicated. If the random vector $\mathbf{X}$ is a random sample, we can generate each independent $X$ using the inverse cdf as above. If it's a multivariate normal $\mathbf{X} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$, we can resolve $\mathbf{X}$ along the eigenvectors of $\boldsymbol{\sigma}^2$ and

Example: Testing the $\chi^2$ for a multinomial (see notebook).

## Thursday, April 24, 2014

## 4.2 Markov Chain Monte Carlo

## 4.3 Problem to be solved

We often want to evaluate multidimensional sums or integrals involving a probability distribution, for instance to calculate an expectation value

$$\langle h(\mathbf{X}) \rangle = \int \cdots \int h(\mathbf{x})\, f(\mathbf{x})\, d^m x \tag{4.1}$$

or a marginalized posterior

$$f(x_1, x_2|D, I) = \int \cdots \int f(x_1, x_2, x_3, \cdots, x_m)\, dx_3 \cdots dx_m \tag{4.2}$$

or even to get the normalization constant when we know that $f(\mathbf{x}) \propto g(\mathbf{x})$, so that

$$f(\mathbf{x}) = \frac{g(\mathbf{x})}{\int \cdots \int g(\mathbf{x}')\, d^m x'} \tag{4.3}$$

If the dimension of the space we're integrating is too high, a traditional numerical sum or integral will be impractical. For instance, if we have 15 parameter values, even just a grid 10 points on a side would have $10^{15}$ points.

Another approach would be to randomly pick points $\mathbf{x}_i$ from a uniform distribution in the space of interest and approximate the expectation value with a weighted average

$$\langle h(\mathbf{X}) \rangle \approx \frac{\sum_i h(\mathbf{x}_i) g(\mathbf{x}_i)}{\sum_i g(\mathbf{x}_i)} \tag{4.4}$$

but for many probability distributions, we'll waste a lot of points in regions where the integrand is small.

The Markov Chain Monte Carlo, or MCMC, method is a way to generate a string of points whose long-term frequency agrees with the probability distribution, so we can evaluate the integrals as sums over points in the chain:

$$\langle h(\mathbf{X}) \rangle \approx \frac{1}{N} \sum_{i=1}^{N} \sum_i h(\mathbf{x}_i) \tag{4.5}$$

We no longer need to weight by the probability density, since the points $\{\mathbf{x}_i\}$ themselves are already distributed according to that density.

### 4.3.1 Metropolis-Hastings method

The so-called Metropolis-Hastings[4] approach follows the following recipe:

1. Pick any point in parameter space to be $\mathbf{x}_0$
2. Repeat the following:
   (a) Randomly generate a proposed new point $\mathbf{y}$ using some probability distribution $q(\mathbf{y}|\mathbf{x}_i)$

[4] Metropolis [*Journal of Chemical Physics* **21**, 1087 (1953)] developed the method under the assumption of a symmetric proposal distribution $q(\mathbf{x}|\mathbf{y}) = q(\mathbf{y}|\mathbf{x})$; Hastings [*Biometrika* **57**, 97 (1970)] put in the extra factors needed to generalize to asymmetric proposal distributions.

(b) Evaluate the probability or probability density at the old an new points, $f(\mathbf{x}_i)$ and $f(\mathbf{y})$.

(c) If $f(\mathbf{y})q(\mathbf{x}_i|\mathbf{y}) \geq f(\mathbf{x}_i)q(\mathbf{y}|\mathbf{x}_i)$, choose $\mathbf{x}_{i+1} = \mathbf{y}$.

(d) If $f(\mathbf{y})q(\mathbf{x}_i|\mathbf{y}) < f(\mathbf{x}_i)q(\mathbf{y}|\mathbf{x}_i)$, make the decision probabilistically:

- With probability $\frac{f(\mathbf{y})q(\mathbf{x}_i|\mathbf{y})}{f(\mathbf{x}_i)q(\mathbf{y}|\mathbf{x}_i)}$, choose $\mathbf{x}_{i+1} = \mathbf{y}$.
- With probability $1 - \frac{f(\mathbf{y})q(\mathbf{x}_i|\mathbf{y})}{f(\mathbf{x}_i)q(\mathbf{y}|\mathbf{x}_i)}$, choose $\mathbf{x}_{i+1} = \mathbf{x}_i$.

Each value in the chain will be correlated with the previous value, and the behavior will therefore be influenced by the starting point. But after an initial "burn-in" period, the chain should settle down into a steady-state where we visit each point with a frequency given by the probability density $f(\mathbf{x})$.

### 4.3.2 MCMC example: multinomial distribution

Gregory gives some examples of MCMC procedures using continuous distributions, but for variety and simplicity I'll consider a discrete distribution, the multinomial distribution, so the pmf we're interested in is

$$p(x_1, \ldots, x_k) = \frac{n!}{x_1! \cdots x_k!} \alpha_1^{x_1} \cdots \alpha_k^{x_k} \tag{4.6}$$

where $\{x_i | i = 1, \ldots, k\}$ are all non-negative integers with $\sum_{i=1}^{k} x_i = n$ and $\{\alpha_i | i = 1, \ldots, k\}$ are known parameters between 0 and 1, with $\sum_{i=1}^{k} \alpha_k = 1$. This is most interesting when $k$ is large, i.e., if we're jumping around a high-dimensional space. However, it's easier to visualize on a lower-dimensional space, so consider the case where $k = 3$, $n = 4$ and $\{\alpha_i\} = \{\frac{1}{2}, \frac{1}{4}, \frac{1}{4}$.

We'll also introduce a very simple proposal distribution: randomly pick two different indices $j$ and $\ell$, and let $y_j = x_j + 1$ and $y_\ell = x_\ell - 1$, with the other elements of $\mathbf{y}$ the same as those of $\mathbf{x}$. For our trinomial case, this just means take a step randomly forwards or back in one of the three possible directions.

We explore this with the ipython notebook `http://ccrg.rit.edu/~whelan/courses/2014_1sp_ASTP_611/data/notes_inference_mcmctrinomial.ipynb`

### 4.3.3 Why the method works

### 4.3.4 Tips and pitfalls

## Tuesday, April 29, 2014

# 5 Estimating Rates from Counting Experiments

*See Gregory, Chapter 14*

A common experiment in Physics and Astronomy involves counting observed events (including, in principle, photons within a spectral channel) and trying to estimate the rate associated with the underlying process. This is often complicated by the presence of *background events* which are not produced by the process in question (as opposed to the *foreground events* we're interested in). Three common scenarios, of increasing complexity are:

1. We observe $k$ events in a time $T$ and want to infer the rate $r$ associated with those events.
2. We know the background rate $b$ and want to infer the foreground (or signal) rate $s = r - b$ from the observation.
3. Both the foreground rate $s$ and the background rate $b$ are unknown, and we make observations both on-source (where the rate will be $s+b$) and off-source (where only background events will be present, and the rate will be $b$).

In all of these cases, the number of events observed should obey a Poisson with a mean equal to the rate times the observation time,

$$p(k|r,I) = \frac{(rT)^k}{k!} e^{-rT} \tag{5.1}$$

We'll mostly consider Bayesian approaches to these problems, but also keep the frequentist prescriptions in mind.

## 5.1 Case 1: No background

### 5.1.1 Frequentist approaches

Frequentist statistics doesn't allow us to define probabilities for the rate $r$ to lie in an interval, but it does allow constructions like the maximum likelihood estimate, which turns out to be

$$\widehat{r} = \frac{k}{T} \tag{5.2}$$

or a confidence interval at confidence level $\alpha$, defined by

$$P(R_\ell \leq r \leq R_u) = \alpha \tag{5.3}$$

where $R_\ell = \ell(K)$ and $R_u = u(K)$ are statistics constructed from the random variable $K$. The measured confidence interval is then $[\ell(k), u(k)]$ For example, if we are simply interested in an upper limit, so that $r_\ell = 0$, we want

$$\alpha = P(r \leq u(K)) = P(u^{-1}(r) \leq K) = \sum_{j=u^{-1}(r)}^{\infty} \frac{(rT)^j}{j!} e^{-rT} \tag{5.4}$$

This looks like a pretty confusing way to define the function $u^{-1}(r)$, but remember, we're interested in $u(k)$ for the actual measured $k$, so it means that if we evaluate (5.4) for $r = u(k)$ (which we can do since it's supposed to be true for any $r$), we get

$$\alpha = \sum_{j=k}^{\infty} \frac{(u(k)T)^j}{j!} e^{-u(k)T} = 1 - \sum_{j=0}^{k-1} \frac{(u(k)T)^j}{j!} e^{-u(k)T} \tag{5.5}$$

which is now an equation which can be solved for any $k > 0$. For example, for $k = 1$ it gives us

$$\alpha = 1 - e^{-u(1)T} \tag{5.6}$$

so

$$u(1) = \frac{\ln \frac{1}{1-\alpha}}{T} \; ; \tag{5.7}$$

for $k = 2$ it says

$$\alpha = 1 - [1 + u(2)T] e^{-u(2)T} \tag{5.8}$$

which is a transcendental equation, but we can solve it numerically for $u(2)T$, given $\alpha$.

### 5.1.2 Bayesian Approach

As usual, the Bayesian approach to the problem is more straightforward; if we want to know about $r$ given that we've seen $k$ events, we just construct a posterior using Bayes's theorem:

$$f(r|k,I) = \frac{p(k|r,I) f(r|I)}{p(k|I)} \propto p(k|r,I) f(r|I) \tag{5.9}$$

The main subtlety is choosing the prior distribution $f(r|I)$. An obvious simple choice is a uniform prior

$$f(r|I_0) = \begin{cases} \frac{1}{r_{\text{max}}} & 0 < r < r_{\text{max}} \\ 0 & \text{otherwise} \end{cases} \tag{5.10}$$

where we will find our calculations simplify greatly if $r_{\text{max}}$ is large enough that $k \ll r_{\text{max}} T$. There are some conceptual problems with a uniform prior, though. For example, if we replaced the rate parameter in question with a scale parameter $\beta = \frac{1}{r}$ we would find that the prior on $\beta$ is no longer uniform, but instead $f(\beta|I_0) \propto \beta^{-2}$.

An alternative is to use the prior

$$f(r|I_1) = \begin{cases} \frac{1}{\ln(r_{\max}/r_{\min})}\,\frac{1}{r} & r_{\min} < r < r_{\max} \\ 0 & \text{otherwise} \end{cases} \tag{5.11}$$

This is often referred to as a Jeffreys prior[5] and you can show that $f(\beta|I_1) \propto \beta^{-1}$. We can also call this "uniform in log rate" because if you do a change of variables to $\lambda = \ln r$ you'll find $f(\lambda|I_1)$ is uniform over the allowed range.

Physically, the $\frac{1}{r}$ prior is appropriate when the rate is uncertain over many orders of magnitude, so e.g., it's as likely to be between $10^{-3}$ Hz and $10^{-2}$ Hz as between $10^{-5}$ Hz and $10^{-4}$ Hz. More likely, we have a sense of what the order of magnitude of the rate should be, so a uniform prior, in addition to being simpler, may actually reflect our knowledge better.

So let's move ahead with the assumption that $p(r|I)$ is constant, so Bayes's theorem tells is that

$$f(r|k,I) \propto p(k|r,I) \propto (rT)^k\, e^{-rT} \tag{5.12}$$

We can get the proportionality constant from normalization, so

$$f(r|k,I) = \frac{(rT)^k\, e^{-rT}}{\int_0^{r_{\max}} (r'T)^k\, e^{-r'T}\, dr'} \tag{5.13}$$

If $k \ll r_{\max}T$, the denominator becomes

$$\int_0^{r_{\max}} (r'T)^k\, e^{-r'T}\, dr' = \frac{1}{T}\int_0^{r_{\max}T} u^k\, e^{-u}\, du \approx \frac{1}{T}\int_0^{\infty} u^k\, e^{-u}\, du = \frac{\Gamma(k+1)}{T} = \frac{k!}{T} \tag{5.14}$$

so

$$f(r|k,I) \approx \begin{cases} \frac{T}{k!}\,(rT)^k\, e^{-rT} & 0 < r < r_{\max} \\ 0 & \text{otherwise} \end{cases} \tag{5.15}$$

Note that this is a Gamma distribution with shape parameter $k+1$ and scale parameter $T$.

## 5.2 Case 2: Known Background

Now we have a case where the actual event rate is the unknown quantity of interest, $s$ (the signal or foreground rate) plus a known background rate $b$, i.e., $r = s + b$. Now, if we knew the exact number of background events, we could subtract that, but as it is, all that's known is the event rate, so there's also randomness in the background, so estimating $s = r - b$ doesn't work out the same as estimating $r$.

### 5.2.1 Frequentist approach and issues

We can proceed mostly as before, for example we have a maximum likelihood estimate of

$$\widehat{s} = \widehat{r} - b = \frac{k}{T} - b \tag{5.16}$$

and likewise our confidence interval could be defined using

$$P(R_\ell - b \leq s \leq R_u - b) = \alpha \tag{5.17}$$

But if we happen to get a small number of events, the results can look weird. For instance, if $k < bT$, the maximum likelihood estimate $\widehat{s}$ would be negative. Similar pathological things can happen with the confidence intervals. This is one of the problems addressed in Feldman and Cousins, "Unified approach to the classical statistical analysis of small signals", *Phys. Rev. D* **57**, 3873 (1998).

[5] This is a slight misnomer, since the Jeffreys prior is defined by a mathematical formula using the likelihood, and for some distributions the uniform prior *is* the Jeffreys prior. To make things more confusing, the Jeffreys prior for the rate parameter in an exponential distribution is proportional to $r^{-1}$ as above, but for a Poisson distribution, it's actually proportional to $r^{-1/2}$.

### 5.2.2 Bayesian method

The construction of the posterior proceeds as before, but now we have

$$f(s|k,I) = \frac{([s+b]T)^k\, e^{-[s+b]T}}{\int_0^{s_{\max}} ([s'+b]T)^k\, e^{-[s'+b]T}\, ds'} = \frac{([s+b]T)^k\, e^{-sT}}{\int_0^{s_{\max}} ([s'+b]T)^k\, e^{-s'T}\, ds'} \tag{5.18}$$

where the constant $e^{-bT}$ cancels out. The denominator can be evaluated as

$$\begin{aligned}\int_0^{s_{\max}} ([s'{+}b]T)^k\, e^{-s'T}\, dr' &= \frac{1}{T}\sum_{j=0}^{k} \frac{k!}{j!(k-j)!} \int_0^{s_{\max}T} u^{k-j}\, (bT)^j\, e^{-u}\, du \\ &\approx \frac{1}{T}\sum_{j=0}^{k} \frac{k!}{j!(k-j)!} \underbrace{\int_0^{\infty} u^{k-j}\, (bT)^j\, e^{-u}\, du}_{\Gamma(k-j+1)=(k-j)!} = \frac{k!}{T}\sum_{j=0}^{k}\frac{(bT)^j}{j!}\end{aligned} \tag{5.19}$$

## Thursday, May 1, 2014

## 5.3 Case 3: Unknown/estimated background

We move now to the general case where the foreground and background rates are both unknown. In order to estimate the foreground rate $s$ and disentangle it from the background rate $b$, we conduct two sets of observations:

- an *OFF-source* observation where the rate of events is $b$, of duration $T_{\text{OFF}}$, in which $k_{\text{OFF}}$ events are observed
- an *ON-source* observation where the rate of events is $s+b$, of duration $T_{\text{ON}}$, in which $k_{\text{ON}}$ events are observed

The probability mass functions associated with the on- and off-source distributions are

$$p(k_{\text{ON}}|s,b,I) = \frac{([s+b]T_{\text{ON}})^{k_{\text{ON}}}}{k_{\text{OFF}}!} e^{-[s+b]T_{\text{ON}}} \tag{5.20}$$

and

$$p(k_{\text{OFF}}|b,I) = \frac{(bT_{\text{OFF}})^{k_{\text{OFF}}}}{k_{\text{OFF}}!} e^{-bT_{\text{OFF}}} \tag{5.21}$$

Our goal is to make an inference about the rate $r$ given the on- and off-source observations; in the Bayesian approach this means working out the posterior pdf $f(r|k_{\text{ON}}, k_{\text{OFF}}, I)$, where the information $I$ includes things like the duration of the observations, but not a specific value for $b$.

### 5.3.1 Qualitative

Roughly speaking, the off-source observation will serve as a sort of calibration and allow us to estimate $b$, albeit with some residual uncertainty. We can then estimate $r$ from the on-source observation, subject to the uncertainty in subtracting the background rate. So the result will look something like

$$b \sim \widehat{b} \pm \delta b \sim \frac{k_{\text{OFF}}}{T_{\text{OFF}}} \pm \frac{\sqrt{k_{\text{OFF}}}}{T_{\text{OFF}}} \tag{5.22}$$

and

$$s \sim \widehat{s} \pm \delta s \sim \frac{k_{\text{ON}}}{T_{\text{ON}}} - \widehat{b} \pm \sqrt{\frac{k_{\text{ON}}}{{T_{\text{ON}}}^2} + (\delta b)^2} \tag{5.23}$$

but this back of the envelope calculation will fail if the numbers of events are small.

### 5.3.2 Bayesian method

We want to work out the posterior pdf $f(s|k_{\text{ON}}, k_{\text{OFF}}, I)$ for the foreground rate, given the on- and off-source observations, which we've marginalized over the background rate $b$. We assume that the priors on the foreground and background rates are uniform, i.e.,

$$f(s|I_0) = \begin{cases} \frac{1}{s_{\max}} & 0 < s < s_{\max} \\ 0 & \text{otherwise} \end{cases} \tag{5.24}$$

and

$$f(b|I_0) = \begin{cases} \frac{1}{b_{\max}} & 0 < b < b_{\max} \\ 0 & \text{otherwise} \end{cases} \tag{5.25}$$

and we'll assume $k_{\text{OFF}} \ll b_{\max} T_{\text{OFF}}$, $k_{\text{ON}} \ll s_{\max} T_{\text{ON}}$ and $k_{\text{ON}} \ll b_{\max} T_{\text{ON}}$ to make the integrals simpler.

There are several equivalent ways to arrive at basically the same expression for the posterior. The first two use the fact that the on- and off-source measurements are independent to work in terms of $p(k_{\text{ON}}, k_{\text{OFF}}|s,b,I) = p(k_{\text{ON}}|s,b,I)\, p(k_{\text{OFF}}|b,I)$.

1. Use Bayes's theorem to get the joint posterior

   $$\begin{aligned} f(s,b|k_{\text{ON}},k_{\text{OFF}},I) &\propto p(k_{\text{ON}},k_{\text{OFF}}|s,b,I)\, f(s,b|I) \\ &= p(k_{\text{ON}},k_{\text{OFF}}|s,b,I)\, f(b|I)\, f(s|I) \end{aligned} \tag{5.26}$$

   and then marginalize over $b$ to get

   $$\begin{aligned} f(s|k_{\text{ON}},k_{\text{OFF}},I) &= \int_0^\infty f(s,b|k_{\text{ON}},k_{\text{OFF}},I) \\ &\propto \int_0^\infty p(k_{\text{ON}}|s,b,I)\, p(k_{\text{OFF}}|b,I)\, f(b|I)\, f(s|I)\, db \end{aligned} \tag{5.27}$$

2. Use Bayes's theorem to write

   $$f(s|k_{\text{ON}},k_{\text{OFF}},I) \propto p(k_{\text{ON}},k_{\text{OFF}}|s,I)\, f(s|I) \tag{5.28}$$

   and get the marginalized likelihood by writing

   $$p(k_{\text{ON}},k_{\text{OFF}}|s,I) = \int_0^\infty p(k_{\text{ON}},k_{\text{OFF}}|s,b,I)\, f(b|I)\, db \tag{5.29}$$

3. Consider $I' = k_{\text{OFF}}, I$ to be the state of information after the off-source experiment, and describe the observation in two steps: First, we get a pdf for $b$ based on the off-source experiment

   $$f(b|I') = f(b|k_{\text{OFF}},I) \propto p(k_{\text{OFF}}|b,I)\, f(b|I) \tag{5.30}$$

   and then use this posterior as the prior on $b$ in the on-source experiment:

   $$\begin{aligned} f(s|k_{\text{ON}},I') &\propto \int_0^\infty p(k_{\text{ON}}|s,b,I')\, f(b|I')\, db\, f(s|I') \\ &\propto \int_0^\infty p(k_{\text{ON}}|s,b,I)\, p(k_{\text{OFF}}|b,I)\, f(b|I)\, db\, f(s|I) \end{aligned} \tag{5.31}$$

   where we use the fact that neither the pmf for the on-source experiment nor the pdf for the signal rate depend on the outcome are directly affected by the results of the off-source experiment, so $p(k_{\text{ON}}|s,b,I') = p(k_{\text{ON}}|s,b,I)$ and $f(s|I') = f(s|I)$.

Of course, it's not surprising that all three approaches give the same expression, since they all just follow the rules of probability. The last approach gives us a head start to constructing the posterior on $s$, since we know the pdf of the background rate after the off-source experiment will be a Gamma distribution

$$f(b|k_{\text{OFF}},I) \propto (bT_{\text{OFF}})^{k_{\text{OFF}}} e^{-bT_{\text{OFF}}} \tag{5.32}$$

Note that this distribution has a mean of $\frac{k_{\text{OFF}}+1}{T_{\text{OFF}}}$ and a width of $\frac{\sqrt{k_{\text{OFF}}+1}}{T_{\text{OFF}}}$, so in the limit of a long off-source observation with many events, we get a more and more sharply-peaked distribution in $b$, which makes the estimation of $s$ tend towards the case of a know background.

Moving on to the construction of the posterior on the foreground rate,

$$
\begin{aligned}
f(s|k_{\text{ON}},k_{\text{OFF}},I) &\propto \int_0^\infty ([s+b]T_{\text{ON}})^{k_{\text{ON}}} e^{-[s+b]T_{\text{ON}}} (bT_{\text{OFF}})^{k_{\text{OFF}}} e^{-bT_{\text{OFF}}}\,db \\
&\propto e^{-sT_{\text{ON}}} \int_0^\infty (s+b)^{k_{\text{ON}}} b^{k_{\text{OFF}}} e^{-b(T_{\text{ON}}+T_{\text{OFF}})}\,db \\
&\propto e^{-sT_{\text{ON}}} \sum_{j=0}^{k_{\text{ON}}} \frac{k_{\text{ON}}!}{(k_{\text{ON}}-j)!j!} (s[T_{\text{ON}}+T_{\text{OFF}}])^j \int_0^\infty u^{k_{\text{ON}}+k_{\text{OFF}}-j} e^{-u}\,du \\
&\propto \sum_{j=0}^{k_{\text{ON}}} \frac{(k_{\text{ON}}+k_{\text{OFF}}-j)!}{(k_{\text{ON}}-j)!j!} \left(1+\frac{T_{\text{OFF}}}{T_{\text{ON}}}\right)^j (sT_{\text{ON}})^j\, e^{-sT_{\text{ON}}} \qquad (5.33)
\end{aligned}
$$

Now, it's pretty straightforward to do the integral over $s$ and work out that normalization constant to get

$$
f(s|k_{\text{ON}},k_{\text{OFF}},I) = \frac{T_{\text{ON}} \sum_{j=0}^{k_{\text{ON}}} \frac{(k_{\text{ON}}+k_{\text{OFF}}-j)!}{(k_{\text{ON}}-j)!j!} \left(1+\frac{T_{\text{OFF}}}{T_{\text{ON}}}\right)^j (sT_{\text{ON}})^j\, e^{-sT_{\text{ON}}}}{\sum_{j'=0}^{k_{\text{ON}}} \frac{(k_{\text{ON}}+k_{\text{OFF}}-j')!}{(k_{\text{OFF}}-j')!} \left(1+\frac{T_{\text{OFF}}}{T_{\text{ON}}}\right)^{j'}} \qquad (5.34)
$$

although in practice the shape of the pdf is more interesting.

**Tuesday, May 6, 2014**

# 6 Choosing a Prior Distribution

In a typical parameter estimation problem, we use some data $D$ to make an inference about one or more parameters, usually denoted by $\theta$. In the Bayesian approach, this is done by using Bayes's theorem to construct the posterior distribution

$$
f(\theta|D,I) \propto p(D|\theta,I) f(\theta|I) \qquad (6.1)
$$

The prior distribution $f(\theta|I)$ is supposed to reflect the plausibility we assign to different values of $\theta$, given any information $I$ we possessed going into the experiment. But sometimes the information $I$ might be difficult to turn into a prior pdf for $\theta$. This week we'll consider a few tricks for constructing useful priors given seemingly incomplete information.

## 6.1 The Jeffreys prior

### 6.1.1 Motivations

The simplest prior one might think of would be something that's uniform over the allowed range of $\theta$ values. This seems pretty clear-cut, but the fact that $f(\theta|I)$ is a probability *density* in $\theta$ causes an issue if we change variables. If, instead of $\theta$, we use a parameter $\tau$ which is some function of $\theta$, the pdf for $\tau$ is

$$
f(\tau) = \frac{dP}{d\tau} = \left|\frac{d\theta}{d\tau}\right| \frac{dP}{d\theta} = \left|\frac{d\theta}{d\tau}\right| f(\theta) \qquad (6.2)
$$

so if $f(\theta)$ is a constant, $f(\tau)$ will in general not be.

As a concrete example, suppose we're considering a Poisson process with an unknown rate $\lambda$. We could start with a uniform prior on $\lambda$, as we did last week, but if we replace the rate parameter $\lambda$ by a scale parameter $\beta = \frac{1}{\lambda}$ (e.g., by re-writing an exponential distribution as a Gamma distribution with $\alpha = 1$), we'd find

$$
f(\beta) = \left|\frac{d\lambda}{d\beta}\right| f(\lambda) \propto \frac{1}{\beta^2} \qquad (6.3)
$$

On the other hand, if $f(\lambda) \propto \frac{1}{\lambda}$ then

$$
f(\beta) \propto \frac{\beta}{\beta^2} = \frac{1}{\beta} \qquad (6.4)
$$

Similarly, if we put a prior $f(\sigma) \propto \frac{1}{\sigma}$ on the standard deviation for a normal distribution, we will also find a prior $f(v) \propto \frac{1}{v}$ on the variance $v = \sigma^2$.

Such a prior, proportional to $\frac{1}{\theta}$, can be thought of as a uniform prior on $\ln\theta$. It's common to refer to this as a "Jeffreys prior" for $\theta$, but that's a slight misnomer, as the general Jeffreys prior has a slightly different definition.

### 6.1.2 Mathematical definition

The Jeffreys prior is a prescription for generating a pdf for a parameter from a likelihood function, in a way which is invariant under reparametrization. It is defined by

$$f(\theta) \propto \sqrt{\mathcal{I}(\theta)} \tag{6.5}$$

where $\mathcal{I}(\theta)$ is the Fisher information, sort of a one-dimensional equivalent of the Fisher information matrix. Given a log-likelihood $\ell(\theta;x) = \ln f(x|\theta)$, we construct

$$-\ell_{,\theta\theta}(\theta;x) = -\frac{\partial^2\ell}{\partial\theta^2} \tag{6.6}$$

and then take the expectation value

$$\mathcal{I}(\theta) = \langle -\ell_{,\theta\theta}(\theta;X)\rangle \tag{6.7}$$

This prescription gives the same expression even if we change variables from $\theta$ to $\tau(\theta)$ in the likelihood, i.e.,

$$f(\tau) \propto \sqrt{\mathcal{I}(\tau)} \tag{6.8}$$

because the partial derivatives arising from the chain rule are just what's needed to take care of the transformation from $f(\theta)$ to $f(\tau)$. It's worth noting that this works because $\ell$ is the log-likelihood, and the likelihood is also used in the expectation value, so

$$\begin{aligned}\mathcal{I}(\theta) &= -\int_{-\infty}^{\infty} \ell_{,\theta\theta}(\theta;x) f(x|\theta)\,dx \\ &= -\frac{\partial}{\partial\theta}\int_{-\infty}^{\infty}\frac{\partial \ln f(x|\theta)}{\partial\theta} f(x|\theta)\,dx + \int_{-\infty}^{\infty}\frac{\partial \ln f(x|\theta)}{\partial\theta}\frac{\partial f(x|\theta)}{\partial\theta}\,dx\end{aligned} \tag{6.9}$$

The first term vanishes because

$$\begin{aligned}\int_{-\infty}^{\infty}\frac{\partial \ln f(x|\theta)}{\partial\theta} f(x|\theta)\,dx &= \int_{-\infty}^{\infty}\frac{1}{f(x|\theta)}\frac{\partial f(x|\theta)}{\partial\theta} f(x|\theta)\,dx \\ &= \int_{-\infty}^{\infty}\frac{\partial f(x|\theta)}{\partial\theta}\,dx = \frac{\partial}{\partial\theta}\int_{-\infty}^{\infty} f(x|\theta)\,dx = \frac{\partial}{\partial\theta}1 = 0\end{aligned} \tag{6.10}$$

and the second term shows us that

$$\mathcal{I}(\tau) = \left(\frac{\partial\tau}{\partial\theta}\right)^2 \mathcal{I}(\theta) \tag{6.11}$$

which gives us just the Jacobian we need to transform $f(\theta)$ into $f(\tau)$.

### 6.1.3 Examples

We can apply this to a few distributions; for an exponential distribution,

$$f(x|\lambda) = \lambda e^{-\lambda x} \tag{6.12}$$

the log-likelihood is

$$\ell(\lambda;x) = \ln\lambda - \lambda x \tag{6.13}$$

so

$$\ell' = \frac{1}{\lambda} - x \tag{6.14}$$

and

$$-\ell'' = \frac{1}{\lambda^2} \tag{6.15}$$

which means $\mathcal{I}(\lambda) = \frac{1}{\lambda^2}$ and the Jeffreys prior is

$$f(\lambda) \propto \sqrt{\mathcal{I}(\lambda)} = \frac{1}{\lambda} \tag{6.16}$$

as expected.

For a Gamma distribution,

$$f(x|\alpha,\beta) = \frac{x^{\alpha-1}}{\Gamma(\alpha)}\beta^{-\alpha}e^{-x/\beta} \tag{6.17}$$

which means

$$\ell = (\alpha-1)\ln x - \ln\Gamma(\alpha) - \alpha\ln\beta - \frac{x}{\beta} \tag{6.18}$$

and

$$\ell_{,\beta} = -\frac{\alpha}{\beta} + \frac{x}{\beta^2} \tag{6.19}$$

and

$$-\ell_{,\beta} = -\frac{\alpha}{\beta^2} + \frac{2x}{\beta^3} \tag{6.20}$$

which makes the Fisher information for the scale parameter $\beta$

$$\mathcal{I}(\beta) = -\frac{\alpha}{\beta^2} + \frac{2\langle X\rangle}{\beta^3} = -\frac{\alpha}{\beta^2} + \frac{2\alpha\beta}{\beta^3} = \frac{\alpha}{\beta^2} \tag{6.21}$$

and the Jeffreys prior

$$f(\beta) \propto \sqrt{\mathcal{I}(\beta)} \propto \frac{1}{\beta} \tag{6.22}$$

For a Gaussian, you can show the Jeffreys prior for $\sigma$ is proportional to $\frac{1}{\sigma}$, but for $\mu$ it is uniform.

There's sort of an odd property to this prescription, though. It requires that you know the likelihood function to construct the prior, i.e., you have to know what experiment you plan to do to construct $f(\theta|I)$. This is really not the sort of information we'd imagine to be part of $I$. To see how this can lead to some weird results, return to the question of a prior for a Poisson event rate $\lambda$, but instead of using the exponential pdf to construct the likelihood function, suppose we plan to count the number of events in a time $T$, and use the Poisson pmf

$$p(k|\lambda,T) = \frac{1}{k!}(\lambda T)^k e^{-\lambda T} \tag{6.23}$$

so the log-likelihood is

$$\ell(\lambda;k) = k\ln\lambda - \lambda T - \ln\frac{T^k}{k!} \tag{6.24}$$

which means

$$\ell_{,\lambda} = \frac{k}{\lambda} - T \tag{6.25}$$

and

$$-\ell_{,\lambda\lambda} = \frac{k}{\lambda^2} \tag{6.26}$$

which makes the Fisher information

$$\mathcal{I}(\lambda) = \frac{\langle K\rangle}{\lambda^2} = \frac{\lambda T}{\lambda^2} \propto \frac{1}{\lambda} \tag{6.27}$$

and the Jeffreys prior

$$f(\lambda) \propto \frac{1}{\sqrt{\lambda}} \tag{6.28}$$

As a curiosity, note that this is the geometric mean of the Jeffreys prior $f(\lambda) \propto \frac{1}{\lambda}$ which we constructed above from the exponential distribution (which is usually what people mean when they talk about the Jeffreys prior for a rate) and the uniform prior we used for counting experiments last week.

## 6.2 Conjugate prior families

Another case where the form of the likelihood function can indicate a convenient choice of prior is that of a conjugate prior family. As an example, consider a counting experiment where as usual

$$p(k|\lambda, T, I) = \frac{1}{k!}(\lambda T)^k e^{-\lambda T} \tag{6.29}$$

and suppose it happens that the prior pdf for $\lambda$ is a Gamma$(\alpha, \beta)$ distribution

$$f(\lambda|I) \propto \lambda^{\alpha-1} e^{-\lambda/\beta} \tag{6.30}$$

Then Bayes's theorem tells us the posterior will be

$$\begin{aligned} f(\lambda|k, T, I) &\propto f(\lambda|I)p(k|\lambda, T, I) \propto \lambda^{\alpha-1}e^{-\lambda/\beta}\lambda^k e^{-\lambda T} \\ &= \lambda^{\alpha+k-1}e^{-\lambda\left(\frac{1}{\beta}+T\right)} \end{aligned} \tag{6.31}$$

which is a Gamma$(\alpha', \beta')$ distribution, where

$$\alpha' = \alpha + k \qquad \text{and} \qquad \frac{1}{\beta'} = \frac{1}{\beta} + T \tag{6.32}$$

I.e., if the prior for the rate in a Poisson experiment is a Gamma distribution, the posterior will also be a Gamma distribution. We say that the Gamma distribution is a *conjugate prior family* for the Poisson distribution. (You saw something similar on the second prelim: if the prior on the mean for a Gaussian experiment is a Gaussian, the posterior will also be a Gaussian.)

Now, again, if the prior $f(\theta|I)$ is supposed to represent your prior knowledge of the parameter $\theta$, there's no reason why it has to be a member of the conjugate prior family. But these families are actually pretty general, so there's often a member which is a good approximation to your knowledge. For example, if you wanted to choose an approximately uniform prior $f(\lambda|I) =$ constant, you could take a Gamma distribution with $\alpha = 1$ and $\beta$ very large. Or if you wanted a prior $f(\lambda|I) \propto \frac{1}{\lambda}$ (the uniform-in-log-$\lambda$ prior which is imprecisely called the Jeffreys prior), you could take $\alpha$ close to zero and $\beta$ very large. (We can't take $\alpha = 0$ because the prior won't be normalizable.) We see from the form of (6.32) that if $\frac{1}{\beta} \ll T$ it doesn't matter what value we actually used for $\beta$, and the results will look like what we got last week with a uniform prior.

["Empirical Bayes" and "the prior gets out of the way".]

# Thursday, May 8 2014

## 6.3 Maximum entropy

*See Gregory, Chapter 8, and Sivia, Chapter 5*

Finally, we consider a prescription for constructing a prior probability distribution when we have incomplete information. The so-called *maximum entropy* prescription says to choose the probability distribution which maximizes the information entropy subject to any constraints. The information entropy[6] is defined as

$$S = -\sum_I p_I \ln p_I \tag{6.33}$$

where $i$ indicates a state of the system, and $p_I$ the probability associated with that state by the probability distribution,

### 6.3.1 Motivation for definition of entropy

Recall that in thermodynamics, the entropy of a state is

$$S = k_B \ln \Omega \tag{6.34}$$

where $k_B$ is Boltzmann's constant, and $\Omega$ is the multiplicity, the number of equivalent ways in which the state can be constructed.

[6] often known as the Shannon information entropy and defined by Claude Shannon in *Bell System Technical Journal* **27,** 379 (1948)

To associate this with a probability distribution, start with the case where there are $K$ different discrete states, each of which has probability $p_I$ according to the probability distribution. The frequency interpretation of probability tells us that if we run the random experiment $N$ times, it will be found in the $I$th state $N_I \approx Np_I$ times. The number of different ways to choose which $N_1$ of the $N$ experiments are in state 1, $N_2$ are in state 2, etc is the

$$\Omega = \frac{N!}{N_1!N_2! \cdots N_k!} \tag{6.35}$$

which makes the entropy, up to a constant,

$$S \propto \ln \Omega = \ln N! - \sum_{I=1}^{K} \ln N_I! \tag{6.36}$$

If $N_I$ is large, Stirling's approximation says

$$\ln N_I! \approx N_I \ln N_I - N_I \tag{6.37}$$

so

$$\ln \Omega \approx N \ln N - N - \sum_{I=1}^{K} N_I \ln N_I + \sum_{I=1}^{K} N_I \tag{6.38}$$

Since $\sum_{I=1}^{K} N_I = N$, two of the terms cancel, and we're left with

$$\ln \Omega \approx \sum_{I=1}^{K} N_I \ln N - \sum_{I=1}^{K} N_I \ln N_I = - \sum_{I=1}^{K} N_I \ln \frac{N_I}{N} = -N \sum_{I=1}^{K} p_I \ln p_I \tag{6.39}$$

since we're only looking to define $S$ up to a constant, we can divide by $N$ to get

$$S \propto \frac{\ln \Omega}{N} = - \sum_{I=1}^{K} p_I \ln p_I \tag{6.40}$$

as advertised.

We'll now show how several familiar distributions arise from maximum entropy arguments in the presence of certain constraints.

### 6.3.2 Example: Uniform Distribution

The minimal constraint which must always be present is the normalization of the probability distribution, that $\sum_I p_I = 1$. Suppose that is the only constraint we need to enforce. We can use the method of Lagrange multipliers and minimize

$$S_{\text{eff}} = - \sum_{I=1}^{K} p_I \ln p_I + \lambda \left( \sum_{I=1}^{K} p_I - 1 \right) \tag{6.41}$$

with respect to $\{p_I\}$ and $\lambda$ to minimize the entropy subject to the constraint, and enforce the constraint itself. Taking $\frac{\partial S_{\text{eff}}}{\partial p_J}$ gives us the equation

$$0 = - \ln p_J - \frac{p_J}{p_J} + \lambda = - \ln p_J - 1 + \lambda \tag{6.42}$$

which says that

$$p_J = e^{\lambda - 1} \tag{6.43}$$

i.e., all of the probabilities are the same. Taking $\frac{\partial}{\partial \lambda}$ tells us that

$$\sum_{I=1}^{K} p_I = 1 \tag{6.44}$$

i.e., $p_I = \frac{1}{K}$, i.e., a uniform distribution.

### 6.3.3 Example: Binomial Distribution

Sometimes it's important to know the nature of the situation described by the probability distribution, and not just think of the pdf or pmf as a function to be derived. For instance, consider the case where the possible outcomes of the experiment are the sequences of answers to $n$ yes or no questions. Then there are $K = 2^n$ possible outcomes, e.g., for $n = 3$ they are $\{YYY, YYN, YNY, YNN, NYY, NYN, NNY, NNN\}$. The calculation just followed says that, in the absence of any constraints, the maximum entropy distribution is for all of the possible outcomes to have the same probability, i.e.,

$$p_I = \frac{1}{2^n} \qquad \text{MaxEnt with no constraints} \tag{6.45}$$

However, we might choose to label the results of the experiment not by the sequence of yes and no answers, but by the total number $k$ of yes results. Then each outcome $I$ has a corresponding $k_I$. The total probability of getting $k$ yes answers is

$$p_k = \sum_{I:k_I=k} p_I \tag{6.46}$$

Let's consider how to construct the entropy from $\{p_k\}$ in a way which will agree with the construction based on $\{p_I\}$. If we restrict attention to probability distributions where all the $\{p_I\}$ corresponding to a given $k$ are equal, and refer to the number of outcomes for a given $k$ as $m_k$, then

$$p_I = \frac{p_{k_I}}{m_{k_I}} \tag{6.47}$$

If we consider that we can write the sum over all outcomes as a sum over $k$ values, where each term in that sum contains a sum over all outcomes for that $k$ value, we have

$$\begin{aligned} S &= -\sum_I p_I \ln p_I = -\sum_k \sum_{I:k_I=k} \frac{p_{k_I}}{m_{k_I}} \ln \frac{p_{k_I}}{m_{k_I}} \\ &= -\sum_k m_k \frac{p_k}{m_k} \ln \frac{p_k}{m_k} = -\sum_k p_k \ln \frac{p_k}{m_k} \end{aligned} \tag{6.48}$$

The $m_k$, which is sort of a multiplicity of states corresponding to the $k$ value, is also known as the *measure* associated with the discrete space of $k$ values.

If, as before, we only require that the probability distribution be normalized ($\sum_{k=0}^{n} p_k = 1$), the entropy is maximized when

$$p_k = m_k \frac{1}{2^n} = \binom{n}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{n-k} \tag{6.49}$$

which we see is the binomial distribution with probability $\frac{1}{2}$.

But what if we add the constraint that $\langle k \rangle = \sum_{k=0}^{n} k\, p_k = \mu$, and find the distribution with maximizes the entropy subject to that constraint? Now there are two constraints, so we include two Lagrange multipliers and find the maximum of

$$S_{\text{eff}} = \sum_{k=0}^{n} p_k \ln \frac{p_k}{m_k} + \lambda_0 \left( \sum_{k=0}^{n} p_k - 1 \right) + \lambda_1 \left( \sum_{k=0}^{n} k p_k - \mu \right) \tag{6.50}$$

Differentiating with respect to one of the $\{p_k\}$ gives

$$0 = \frac{\partial S_{\text{eff}}}{\partial p_k} = \ln \frac{p_k}{m_k} + \frac{p_k}{p_k} + \lambda_0 + k\lambda_1 = \ln \frac{p_k}{m_k} + 1 + \lambda_0 + k\lambda_1 \tag{6.51}$$

which means

$$p_k = m_k e^{1+\lambda_0+k\lambda_1} = \binom{n}{k} e^{1+\lambda_0} \left(e^{\lambda_1}\right)^k \tag{6.52}$$

Setting the derivatives $\frac{\partial S_{\text{eff}}}{\partial \lambda_0}$ and $\frac{\partial S_{\text{eff}}}{\partial \lambda_1}$ to zero gives us the two constraints, so we need to choose $\lambda_0$ and $\lambda_1$ to ensure $\sum_{k=0}^{n} p_k = 1$ and $\sum_{k=0}^{n} k\, p_k = \mu$. Now, we already know a distribution of the form (6.52) which satisfies the constraints, although we may not recognize it in that form. Consider the binomial distribution with $n$ trials and a probability of $\mu/n$. It has

$$p_k = \binom{n}{k} \left(\frac{\mu}{n}\right)^k \left(1 - \frac{\mu}{n}\right)^{n-k} = \binom{n}{k} \left(1 - \frac{\mu}{n}\right)^n \left(\frac{\mu/n}{1-\mu/n}\right)^k \tag{6.53}$$

so evidently, if we set

$$e^{1+\lambda_0} = \left(1 - \frac{\mu}{n}\right)^n \tag{6.54}$$

and

$$e^{\lambda_1} = \frac{\mu/n}{1-\mu/n} \tag{6.55}$$

the distribution (6.52) will satisfy the constraints. Thus the maximum entropy distribution for the case where the expectation value of the number of yes results in a sequence is known is the binomial distribution.

Note that neither in this case nor in the unconstrained case did we assume the yes/no questions were associated with repeated identical trials with the same probability of success. I.e., we need not be flipping the same coin; it could be different coins. On the homework you'll consider the scenario where we are doing repeated trials with the same unknown probability, and see the implications of choosing the maximum entropy distribution for that probability.

### 6.3.4 Continuous distributions

Many probability distributions are continuous rather than discrete. The natural thing is to replace the sum weighted by the probability distribution $p_k$ with an integral weighted by the probability density function $f(x)$. If we also replace the measure $m_k$ with something which acts like a density in $x$, then we can define an expression

$$S = -\int_{-\infty}^{\infty} f(x) \ln \frac{f(x)}{m(x)}\, dx \tag{6.56}$$

which is invariant under a change of variables from e.g., $x$ to $y(x)$, since $f(x)\,dx = f(y)\,dy$, $m(x)\,dx = m(y)\,dy$, and

$$\frac{f_X(x)}{m_X(x)} = \frac{f_Y(y(x))}{m_Y(y(x))} \tag{6.57}$$

We call $m(x)$ the Lebesgue measure, and it's whatever is natural for the variable(s) in question. For instance, since $dx\,dy = r\,dr\,d\phi$, $m(x,y) = 1$ and $m(r,\phi) = r$.

The other added complication is that varying $S$ with respect to $f(x)$ is a functional derivative, but it generally works out.

As an example, if we assume $m(x) = 1$, $-\infty < x < \infty$, we can show that the maximum entropy distribution with expectation value $\mu$ and variance $\sigma^2$ is a Gaussian. We do this by varying

$$\begin{aligned} S_{\text{eff}} = &\int_{-\infty}^{\infty} f(x) \ln f(x)\, dx + \lambda_0 \left(\int_{-\infty}^{\infty} f(x)\, dx - 1\right) \\ &+ \lambda_1 \left(\int_{-\infty}^{\infty} x f(x)\, dx - \mu\right) + \lambda_2 \left(\int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx - \sigma^2\right) \end{aligned} \tag{6.58}$$

with respect to $f(x)$, $\lambda_0$, $\lambda_1$, and $\lambda_2$. **Exercise:** try this!